Mining Political Blogs With Network Based Topic Models

by

Jiawei Liang

Department of Statistical Science and Department of Economics
Duke University

Date: _____

Approved:

_____
David L. Banks, Supervisor

_____
David B. Dunson

_____
Michael C. Munger

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science and Department of Economics
in the Graduate School of Duke University

2014

Abstract

Mining Political Blogs With Network Based Topic Models

by

Jiawei Liang

Department of Statistical Science and Department of Economics
Duke University

Date: _____

Approved:

_____
David L. Banks, Supervisor

_____
David B. Dunson

_____
Michael C. Munger

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science and Department of Economics
in the Graduate School of Duke University

2014

Abstract

We develop a Network Based Topic Model (NBTM), which integrates a Random Graph model with the Latent Dirichlet Allocation (LDA) model. The NBTM assumes that the topic proportion of a document has a fixed variance across the document corpus with author differences treated as random effects. It also assumes that the links between documents are binary variables whose probabilities depend upon the author random effects. We fit the model to political blog posts during the calendar year 2012 that mention Trayvon Martin. This paper presents the topic extraction results and posterior prediction results for hidden links within the blogosphere.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

K	Number of topics
α	Bayesian hyperparameter for topic proportions
ϕ	Bayesian hyperparameter for word proportions
η	Bayesian hyperparameter for blog links
θ_d	Topic proportions of document d
$\theta_{d,b}$	Topic proportions of document d belonging to blog domain b
β_k	Word proportions of topic k
$w_{d,i}$	The ith word in document d
$z_{d,i}$	Topic assignment of word $w_{d,i}$
γ_b	Random effect of blog domain b on topic proportion $\theta_{d,b}$
$y_{b,b'}$	Binary variable indicating directed link of blog b to blog b'
Y	Contingency table of random graphs
$n_{d,k}^{(-d,b,i)}$	Count of $z_{d,b,i} = k$ in document d excluding $z_{d,b,i}$
$n_{k,w}^{(-d,i)}$	Word count of $w_{d,i}$ in topic k among all documents excluding the word itself

Abbreviations

tf-idf	Term frequency-inverse document frequency
LSI	Latent Semantic Indexing
pLSI	Probabilistic Latent Semantic Indexing

LDA	Latent Dirichlet Allocation
AT	Author-Topic
TAP	Topic Affinity Propagation
NBTM	Network Based Topic Model
EM	Expectation-Maximization
MCMC	Markov Chain Monte Carlo

Acknowledgements

The dataset used in this study was contributed by Maxpoint Interactive Inc.

1

Introduction

In the past few decades, the growth of text information, such as blogs, news, research articles and the Wikipedia, has generated a new research focus in data mining and machine learning: Topic Modeling.

There were some critical points in the history of topic modeling. One possible starting point for its development is the term frequency-inverse document frequency (tf-idf) scheme (Salton and McGill, 1983), which builds a basic vocabulary for text corpus and stores information about natural language words through a word frequency count matrix. This scheme was further examined by Deerwester et al. (1990), which proposed the idea of latent semantic indexing (LSI) to enable spectral analysis of these matrices for dimension compression. To study the statistical relationship within and between documents, Hofmann (1999) proposed a probabilistic model, known as probabilistic LSI (pLSI), which assumed that words in documents are generated from different topics and that a document consists of mixture of topics. Such a model, however, does not contain a probabilistic generative process for the vectors of topic proportions for documents. To overcome this defect and enable Bayesian modeling techniques, Blei et al. (2003) introduced Latent Dirichlet Alloca-

tion (LDA) to extend the pLSI model with an additional generative process for topic proportions through a hierarchical model.

The Blei et al. (2003) approach is a useful technique to study the distribution of words in documents and a considerable number of research papers have been based on this approach. To allow topics to evolve dynamically, Blei and Lafferty (2006) uses the Chinese Restaurant Process and Wang and McCallum (2006) assumes topic proportions change slowly over time. Blei and Lafferty (2009) uses recursive hypothesis tests to bundle words generated from the same topic which lie in proximate positions in a document. Some researchers, including Hofman and Wiggins (2008), Airoldi et al. (2008) and Chang and Blei (2009) examine networks of documents that allow links between documents to be related to the generative process for the words in ways that predict connections among documents.

However, none of these approaches easily incorportes observed node attributes. For example, several articles written by a particular researcher in a short time span probably contain very similar topics, and two bloggers that write on the same subject are likely to link to each other. Therefore, inclusion of network effects in topic modeling might increase the accuracy of word prediction and link prediction.

From a different perspective, some researchers study interaction effects in social networks. Rosen-Zvi et al. (2004) extended LDA to the Author-Topic (AT) model to include authorship information in the model. McCallum et al. (2005) further included linkage information in the AT model to uncover people's roles in social network models. Tang et al. (2009) assumed people's interests were latent topics in a social network model and proposed Topic Affinity Propagation (TAP) to quantify a person's influence in large social networks. However, such models are based on the assumptions that a set of recipients is observed, and that the distribution of topic assignments does not depend on the network but only on the author and the recipient (for directed edges). One might easily come up with an ordinary situation

contradicting the assumption: the recipients of a national newspaper are the people in the country, which is difficult to study. Similarly, the study of retweeting in Twitter might be based upon hashtags, with the original author and potential viewers taken as a network.

In this study, we propose a model integrating LDA and a random graph model, which we call the Network Based Topic Model (NBTM). We use NBTM to analyze documents that are political blogs, which is an example of the situation described above since bloggers provide hyperlinks to articles with similar or opposing political viewpoints, or to articles that review a political issue or event.

The remainder of the paper is organized as follows: Chapter 2 describes the source and quality of the dataset. Chapter 3 introduces the structure of NBTM and its inferential strategy. Chapter 4 highlights the experimental results of topic extraction and hidden link prediction. Finally, Chapter 5 draws conclusions and states possible work for further improvement.

2

Data

The shooting of Trayvon Martin was a major political focus in 2012 (for details, see Appendix A). We use the dataset described in Soriano et al. (2013), who first obtained a list of 1509 top-ranked political blog sites at Technorati[1]. Then MaxPoint Interactive Inc. provided assistance in scraping and tokenizing all blog posts that appeared at those sites during the calendar year of 2012. Of these blog entries, 1103 came from 145 unique domains and contained the unique keyword "Trayvon Martin". These posts formed the text corpus. In order to improve data quality, a further manual examination was performed to assess accuracy and confirm the dates of the blog entries posted. Each domain was labeled as "liberal", "conservative" or "moderate" based on information from the website Technorati. Figure 2.1 shows the document counts by political ideology during the 2012 time span.

From Figure 2.1 we can observe that the number of liberal blog entries and conservative blog entries are about equal. Most of the blogging activities occurred before May 2012 with sporadic peaks responding to the release of new details about the shooting or comments by prominent politicians or legal motions in the Zimmerman

[1] www.techorati.com

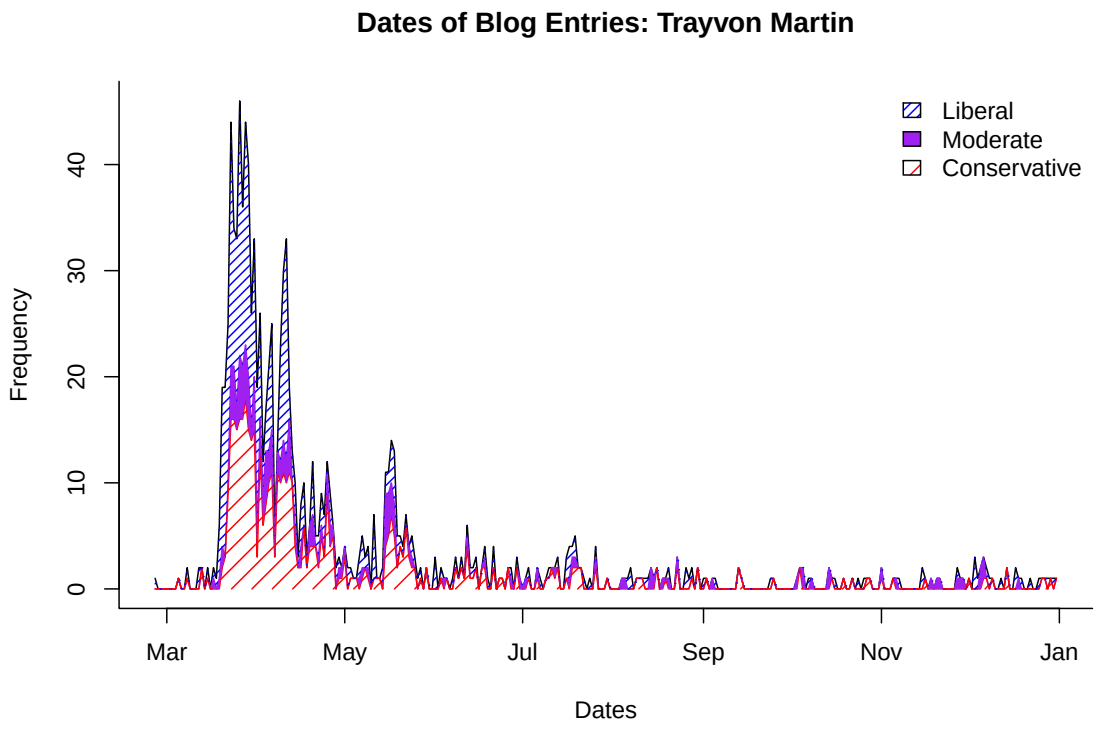

FIGURE 2.1: The frequency of blog entries by political ideology from February 26, 2012 to December 31, 2012. Shading between lines represents frequencies of the blogs of the corresponding political ideology posted on corresponding dates.

trial.

Although there exist errors (e.g. a document dated before the Trayvon Martin incident on February 26, 2012), such noise effects are largely offset by high quality of the remaining dataset and the substantial quantity of material. The network structure in the data arises because each post in the dataset has none to several referencing links to other blog sites.

The case of Travyon Martin is suitable for our study because of its wide coverage and national attention, and also because of the multiple topics that were involved in the debate. Moreover, some topics are highly polarizing, so bloggers form tight networks with those who agree with them, and argue with other cliques with whom

they disagree (Munger, 2008). Due to this segmentation, links within political blogs may be quite informative regarding similarity of topic and stimulation of additional blog posts over time. For the same reasons, normal communication between political groups is probably limited, and the existing communicating links are more likely to be "vituperation than argument" (Munger, 2008).

3

The Approach

3.1 Latent Dirichlet Allocation

The Network Based Topic Model (NBTM) integrates a word-document based topic model and a domain-based random graph model. First we review the Latent Dirichlet Allocation model, then the network model.

The basic latent Dirichlet allocation (LDA) (Blei et al., 2003) model is a hierarchical probabilistic generative model. The model assumes that the text corpus consist of several documents. Each document contains a collection of words, and each word of the document has a topic assignment determined by a vector of topic proportions. Each word in the document is drawn from the distribution of words proportion corresponding to the specific topic assignment. Statistically, the generative process of the model can be written as follows:

1. A fixed number of topics K is chosen, and α and ϕ are positive parameter vectors;

2. The topic proportions of document d can be generated by

$$\theta_d \sim Dirichlet(\alpha) \tag{3.1}$$

3. The word distribution of kth topic can be generated by

$$\beta_k \sim Dirichlet(\phi) \tag{3.2}$$

4. The topic assignment $z_{d,i}$ of word w_i in document d can be generated by

$$z_{d,i} \sim Multinomial(\theta_d) \tag{3.3}$$

5. Finally, word $w_{d,i}$ is generated by the word distribution of its topic assignment:

$$w_{d,i} \sim Multinomial(\beta_{z_{d,i}}) \tag{3.4}$$

The graphical structure of LDA is presented in Figure 3.1 below.

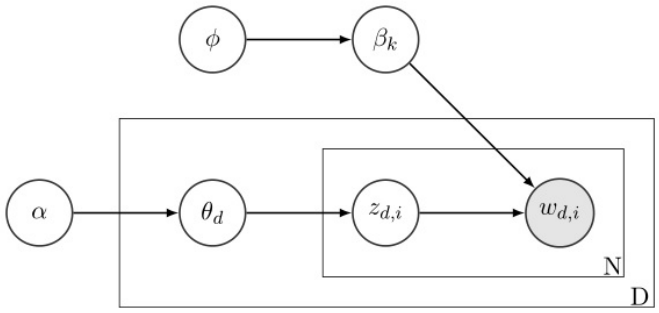

FIGURE 3.1: The graphical structure of LDA. White circles represents latent variables and the shaded circle represents observed variable. Plates represents replications.

While the structure of the model is simple, its likelihood function is intractable (Blei et al., 2003). To make estimates, a collapsed Gibbs sampling algorithm was proposed by Teh et al. (2006) and an expectation-maximization (EM) algorithm for LDA (Blei et al., 2003) is also used.

3.2 Random Graph Model

In the simplest case, consider the Erdős-Rényi-Gilbert Random Graph Model, in which all (undirected) edges between nodes are independent with probability p (Erdős and Rényi, 1960). In this case, if the adjacency matrix is observed (and observing the adjacency matrix is equivalent to observing the network), then the likelihood of the model can be calculated by (Goldenberg et al., 2010)

$$\ell(Y|p) = \prod_{i \neq j} p^{Y_{ij}}(1-p)^{1-Y_{ij}} \tag{3.5}$$

However, as a minor generalization, the probability of an edge might depend on node attributes. We can extend the model as assuming p to be a function that depends on characteristics of node i and j. Namely,

$$Y_{ij}|p \sim Bernoulli(p(i,j)). \tag{3.6}$$

3.3 Network Based Topic Model

Integrating basic latent Dirichlet allocation model and the random graph model, we now extend LDA into a network-based topic model.

3.3.1 Modeling assumptions

We assume that the topic proportion of document d depends not only on a fixed effect α, but also depends on a random effect of γ_b that is unique to blog domain b. In addition to affecting topic proportions, the random effects γ_b also serve as factors determining the probabilities of links between two blog domains b and b', denoted by $y_{b,b'}$. The remaining assumptions are similar to those of the basic LDA model. The probabilistic generative process of the model can be described as:

1. Choose a fixed number to total topics K;

2. For each document which belongs to blog domain b, the topic proportion is generated by:

$$\theta_{d,b}|\alpha,\gamma_b \sim Dirichlet(\alpha+\gamma_b) \tag{3.7}$$

3. The topic assignment of word w_i in document d can be generated by

$$z_{d,b,i}|\theta_{d,b} \sim Multinomial(\theta_{d,b}) \tag{3.8}$$

4. The word proportion of topic k, β_k can be generated by

$$\beta_k \sim Dirichlet(\phi) \tag{3.9}$$

5. Then word $w_{d,i}$ is generated by the word distribution of its topic assignment:

$$w_{d,i}|z_{d,b,i},\beta_K \sim Multinomial(\beta_{z_{d,b,i}}) \tag{3.10}$$

6. A binary indicator of links between two blog domains b and b' is then drawn from

$$y_{b,b'}=1|\gamma_b,\gamma_{b'} \sim logit(a(v-\frac{\gamma_b^T\gamma_{b'}}{\|\gamma_b\|})) \tag{3.11}$$

where $\eta=(a,v)$ is the parameter that normalize the probability and control the rate of effect of projection $\gamma_b^T\gamma_{b'}/\|\gamma_b\|$ to the binary indicator.

Figure 3.2 presents the graphical structure of its generative process.

3.3.2 Inference and estimation

A major interest of inference in this problem is to estimate the latent variables plotted in Figure 3.2. The likelihood function of this model, however, is also intractable, so a Markov Chain Monte Carlo (MCMC) updating scheme is implemented. The full joint distribution of the model can be written as

$$L=\prod_d P(\theta_{d,b}|\alpha,\gamma_b)\Big(\prod_i P(z_{d,b,i}|\theta_{d,b})P(w_{d,i}|z_{d,b,i},\beta_k)\Big)\prod_{b\neq b'}P(y_{b,b'}|\gamma_b,\gamma_{b'},\eta) \tag{3.12}$$

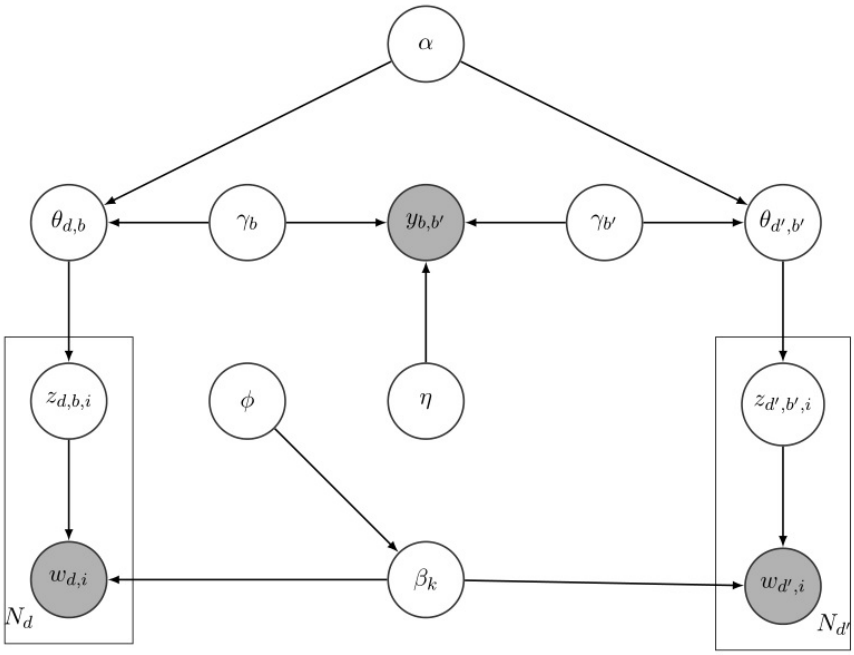

FIGURE 3.2: The graphical structure of the network-based hierarchical topic model. White circles represents latent variables and the shaded circle represents the observed variable. The binary variable $y_{b,b'}$ indicates whether blog domain b and b' are linked. Plates represents replication.

We observe that conjugacy remains for the variables θ, z and β, as in basic LDA model, and only the updating scheme of γ_b is difficult to calculate. While the full conditional distribution of γ_b is still intractable, an approximate updating scheme is provided by Hoff et al. (2002). Thus the MCMC updating scheme can be written as follows:

1. Update the topic proportions of each document belonging to blog domain b by

$$\theta_{d,b} \sim Dirichlet(\alpha + \gamma_b + \sum_i z_{d,b,i}) \tag{3.13}$$

2. Update the word proportions in each topic by

$$\beta_k \sim Dirichlet(\beta + \sum_i w_{d,b,i}) \tag{3.14}$$

3. Update the topic assignment of each word in document d by

$$z_{d,b,i} = k \propto (n_{d,k}^{(-d,b,i)} + \alpha_k + \gamma_{b,k}) \frac{n_{k,w}^{(-d,i)} + \phi_w}{n_k + \sum_w \phi_w - 1} \tag{3.15}$$

where $n_{d,k}^{(-d,b,i)}$ denotes the count of $z_{d,b,i} = k$ in document d, excluding $z_{d,b,i}$, and $n_{k,w}^{(-d,i)}$ denotes the word count of $w_{d,i}$ in topic k in all documents, excluding the word itself.

4. Update γ_b as in Hoff et al. (2002)

4

Experimental Results

4.1 Topic Model

We fit the political blog dataset described in Chapter 2 using the Network Based Topic Model. We first perform the tokenizing transformation as mentioned in Soriano et al. (2013). We assume that each domain has a single node attribute and referencing and being referenced are directed edges in the graph. By using the assumption that there are no time-dependent variables in the model, we collapse multiple edges to single edge, which means the links of domain b referencing domain b' on multiple dates are counted as once. (Note that this causes information loss.)

While we might suspect that both topic proportions and probabilities of occurrence of a single word in a topic might vary over time, political communications are often assumed to be strategic (Manheim, 1991) and, as such, might be invariant over short time periods of time in a stable media environment. Although multiple connections in a certain time period are collapsed to single link and thus we lose connection strength signals, the fundamental connection between two political blogs is arguably invariant during the time interval in this study.

Table 4.1: Results of fitting our model to all documents in the dataset. The 15 words with highest probabilities in each topic are listed.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
"obama"	"like"	"obama"	"law"	"georg zimmerman"
"comment"	"would"	"februari"	"gun"	"trayvon martin"
"american"	"think"	"presid"	"stand ground"	"polic"
"presid"	"dont"	"murder"	"alec"	"said"
"nation"	"peopl"	"news"	"forc"	"call"
"govern"	"one"	"chri"	"state"	"comment"
"left"	"know"	"barack"	"shoot"	"case"
"year"	"even"	"new"	"self defens"	"shot"
"new"	"black"	"women"	"legisl"	"report"
"conserv"	"get"	"presidenti"	"use"	"investig"
"call"	"make"	"elect"	"florida"	"evid"
"america"	"march"	"janet"	"kill"	"time"
"countri"	"thing"	"gop"	"arm"	"arrest"
"polit"	"see"	"read"	"weapon"	"florida"
"state"	"say"	"investor"	"deadli"	"attorney"

After tokenizing the keywords that are probably connected in this context (e.g. the words "trayvon", "martin" and "trayvon martin" should be considered the same word in the corpus) (Soriano et al., 2013), we have fit $K = 5$ topics for all documents and we list the top 15 words with highest probability in each topic in Table 4.1.

According to top-ranked words in each topic, we can assign a practical topic based on personal experience. Topic 2 consists of most common words in English writing and can be considered as noise effect (i.e., blog-speak). Topics 1, 3, 4 and 5 can be interpreted as American politics, the presidential election, "Stand Your Ground" laws, and shooting case details. Comparing Table 4.1 to the result in Soriano et al. (2013), the topic of racism is not significant in this model.

In order to examine how topics vary through the whole time span, we manually selected three break points: March 23, 2012 (when President Barack Obama addressed about the case in public for the first time), April 11, 2012 (when Zimmer-

Table 4.2: Results of fitting our model to documents before March 22, to documents between March 23 to April 10, to documents between April 11 to May 7, and to documents between May 8 to December 31, 2012. The 6 words with highest probabilities in each topic are shown in the table.

Time	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Before March 22, 2012	"said"	"holder"	"law"	"comment"	"georg zimmerman"
	"would"	"new"	"florida"	"read"	"trayvon martin"
	"one"	"american"	"forc"	"get"	"polic"
	"peopl"	"obama"	"stand ground"	"media"	"case"
	"right"	"ralli"	"defend"	"might"	"mar"
	"say"	"republican"	"prosecut"	"fox"	"call"
March 23, 2012– April 10, 2012	"black"	"georg zimmerman"	"state"	"like"	"georg zimmerman"
	"obama"	"trayvon martin"	"law"	"peopl"	"voic"
	"white"	"polic"	"govern"	"one"	"expert"
	"presid"	"said"	"libertarian"	"think"	"nbc"
	"said"	"law"	"year"	"dont"	"owen"
	"racial"	"gun"	"group"	"get"	"audio"
April 11, 2012– May 7, 2012	"obama"	"april"	"comment"	"way"	"georg zimmerman"
	"american"	"repli"	"would"	"wed"	"trayvon martin"
	"state"	"log"	"like"	"stori"	"case"
	"alec"	"regist"	"one"	"news"	"said"
	"presid"	"rick"	"peopl"	"citi"	"charg"
	"right"	"imax"	"black"	"war"	"law"
May 8, 2012– December 31, 2012	"like"	"obama"	"obama"	"would"	"georg zimmerman"
	"one"	"american"	"februari"	"peopl"	"trayvon martin"
	"think"	"state"	"murder"	"say"	"gun"
	"even"	"year"	"presid"	"dont"	"case"
	"black"	"govern"	"elect"	"get"	"kill"
	"august"	"presid"	"women"	"right"	"decemb"

man was charged with second-degree murder), and May 8, 2012 (when Zimmerman's written plea of not guilty was accepted). We study the documents between these break points and, as before, fit the proposed model for $K = 5$ topics. The result of top 6 words in each topics within each time span are shown in Table 4.2.

Comparing Table 4.1 and Table 4.2, the topics in the political blog posts changed through time. While details of the case and blog-speak are included in all periods, the self-defense law is only significant in the first two periods. In addition, in the

second period, two topics containing different details of the case are obtained (Topic 2 and Topic 5). Racism is significant in the second period (Topic 1), and can also be seen in the third period (Topic 3). In the last period, only the topic of case details is remains in documents.

4.2 Predicting Hidden Links

Another major advantage of using the NBTM is to involve text information in link prediction. The collapsed graph of the whole set of documents used in this study is shown in Figure 4.1.

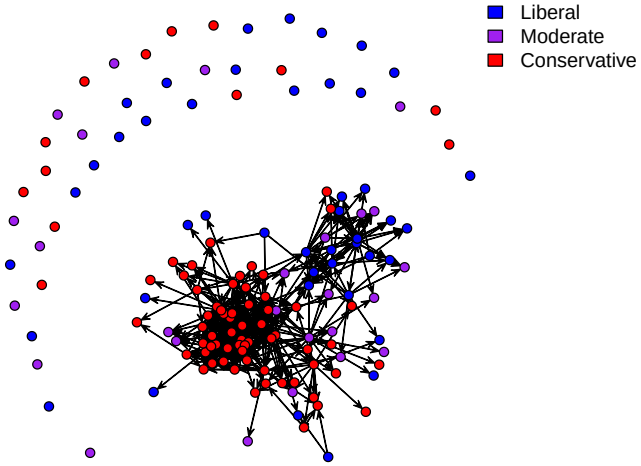

FIGURE 4.1: A social network directed graph demonstration of structures in the blog domains. Blue, purple, and red points represent domains labeled as liberal, moderate and conservative, respectively. Small arrows indicate which site is referencing the other site: the arrow points to the site being referenced.

In Figure 4.1, there are three observable blocks: conservatives tend to group together and form links with each other, while liberals standing at the other end of

the graph, with moderate blogs placed between them. This is not surprising, but it confirms our expectations.

Implementing MCMC algorithm, we use the MCMC samples to conduct posterior prediction of linkages. Using samples of γ_b and η, we can generate posterior predictive samples of $y_{b,b'}$ by Equation 3.11. Using the following prediction criteria, we can assess the hidden links in this social network:

$$Y_{b,b'} = \begin{cases} 1 & \text{if } \overline{y_{b,b'}} \geqslant 0.7 \\ 0 & \text{if } \overline{y_{b,b'}} < 0.7 \end{cases} \tag{4.1}$$

We implement the prediction scheme with the four partitions of documents as described in Section 4.1. The original network of blog domains and the predicted network of blog domains in each time period are shown in Figure 4.2. Red arrows in the network indicate that the link in the predicted network is unobserved in the original network.

By manual observation, the predictive graph and the original graph share similar structure. The center of all four networks is a small group of conservatives with liberal blogs in the distance, which might indicate and verify the assumption of political polarity.

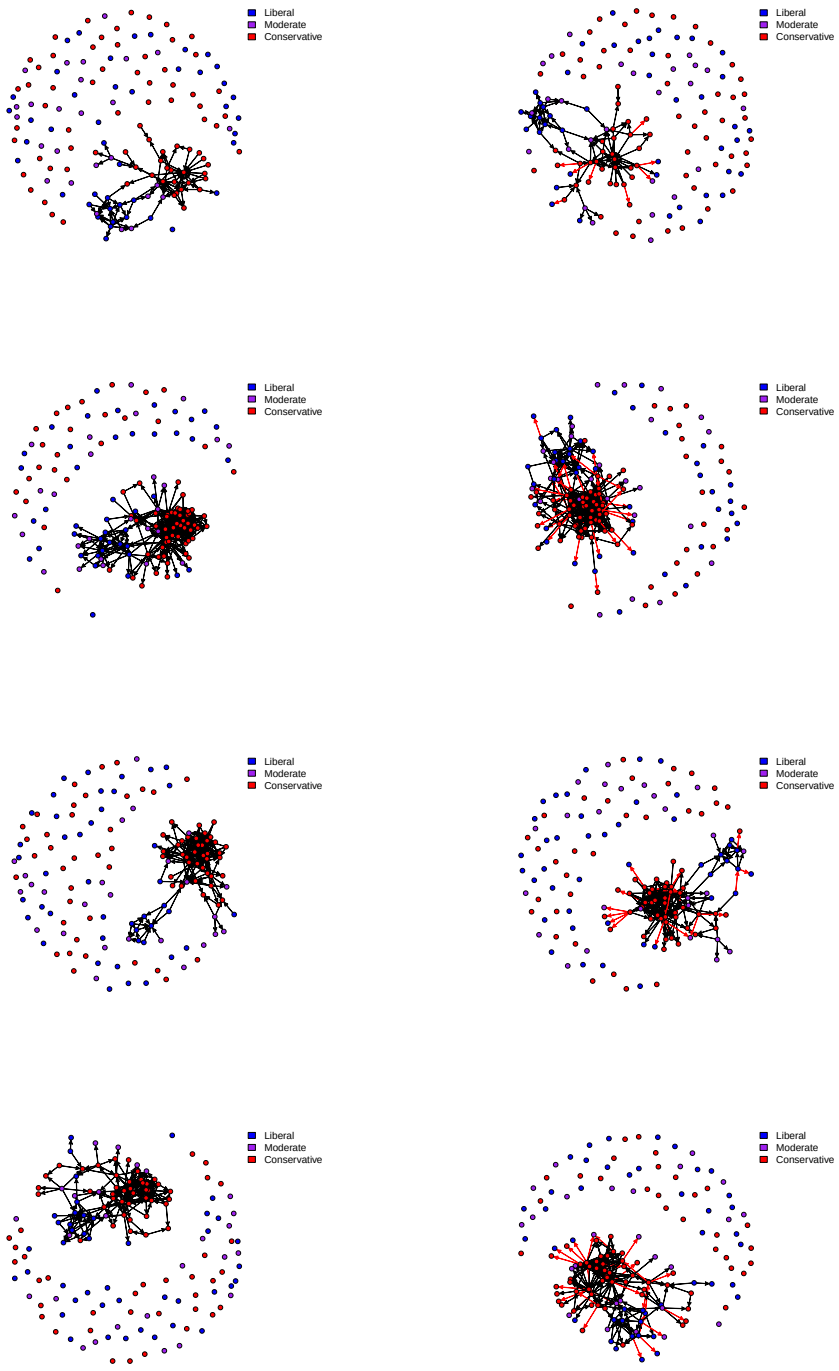

FIGURE 4.2: Social network posterior predictive graphs for different sets of documents. Left: original network of documents. Right: predictive network of the corresponding set. From top to bottom: time periods before Mar. 22, between Mar 23 to Apr. 10, between Apr. 11 to May 7, between May 8 to Dec. 31, 2012.

5

Conclusion

This study proposes the Network Based Topic Model, which integrates a random graph model with the basic Latent Dirichlet Allocation model. This enables us to model blog domain effects as random effects, and use topic distribution inferences on documents to improve link prediction and link prediction to improve topic discovery. In addition, the study used posterior predictive samples to predict links in the political blogosphere.

The model can be applied to any text corpus containing a relatively small number of authors and a large number of documents. In the Trayvon Martin case study, the model appears to have extracted useful topic and network information.

A major disadvantage of the model is that it lost information by collapsing multiple links between sites over time into a single link between those sites. A plausible further improvement of the model is to turn link counts into a measure of connection strength. Future work on the model might also relax the single author assumption and extend the treatment of variation over time.

Appendix A

Background on the Travyon Martin Incident

On the night of February 26, 2012, George Zimmerman (a Hispanic-American) observed Travyon Martin (an African-American) returning home at The Retreat at Twin Lakes, a multi-racial gated community in Sanford, Florida. (Robles, 2012) Zimmerman then called the Sanford police non-emergency number to report Martin as a suspicious person. According to the audio tape of the call, Zimmerman was following Martin and then lost sight of him. (Robles, 2012) Soon after the call ended, there was a violent encounter between Zimmerman and Martin which ended with Zimmerman fatally shooting Martin. Zimmerman later claimed the shooting was in self-defense.

The case was covered only in local media for the first ten days before the Martin family attorney sought attention from national media. The coverage and debate continued to grow as more details, including audio tapes and witness accounts, were released to the public. Besides wide coverage of the facts of the case, it also prompted discussion of social issues such as racism and self-defense laws, including Florida's "stand your ground" law and the "castle doctrine".

A number of key events drove the blog discussion of the Martin case. On March 23, 2012, President Barack Obama addressed the case in public for the first time. On April 11, 2012, George Zimmerman was arrested and charged with second-degree murder. On May 8, 2012, Zimmerman entered a written plea of not guilty. Eventually, in July 2013, Zimmerman was found not guilty (Alcindor, 2012).

Bibliography

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research*, 9, 3.

Alcindor, Y. (2012), "George Zimmerman found not guilty," http://www.usatoday.com/story/news/nation/2013/07/13/george-zimmerman-found-not-guilty/2514163/ (accessed March 2014).

Blei, D. M. and Lafferty, J. D. (2006), "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine Learning*, pp. 113–120, ACM.

Blei, D. M. and Lafferty, J. D. (2009), "Visualizing topics with multi-word expressions," *arXiv preprint arXiv:0907.1013*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *the Journal of Machine Learning Research*, 3, 993–1022.

Chang, J. and Blei, D. M. (2009), "Relational topic models for document networks," in *International Conference on Artificial Intelligence and Statistics*, pp. 81–88.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990), "Indexing by Latent Semantic Analysis," *JASIS*, 41, 391–407.

Erdős, P. and Rényi, A. (1960), "On the evolution of random graphs," *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5, 17–61.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A survey of statistical network models," *Foundations and Trends in Machine Learning*, 2, 129–233.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the american Statistical association*, 97, 1090–1098.

Hofman, J. M. and Wiggins, C. H. (2008), "Bayesian approach to network modularity," *Physical review letters*, 100, 258701.

Hofmann, T. (1999), "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, ACM.

Manheim, J. B. (1991), *All of the people, all the time: Strategic communication and American politics*, ME Sharpe.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005), "Topic and role discovery in social networks," *Computer Science Department Faculty Publication Series*, p. 3.

Munger, M. C. (2008), "Blogging and political information: truth or truthiness?" *Public Choice*, 134, 125–138.

Robles, F. (2012), "A look at what happened the night Trayvon Martin died," http://www.tampabay.com/news/publicsafety/crime/a-look-at-what-happened-the-night-trayvon-martin-died/1223083 (accessed March 2014).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004), "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494, AUAI Press.

Salton, G. and McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill.

Soriano, J., Au, T., and Banks, D. (2013), "Text mining in computational advertising," *Statistical Analysis and Data Mining*, 6, 273–285.

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009), "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 807–816, ACM.

Teh, Y. W., Newman, D., and Welling, M. (2006), "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in *NIPS*, vol. 6, pp. 1378–1385.

Wang, X. and McCallum, A. (2006), "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433, ACM.